MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning

Year: 2023 | *Citations: 602* | *Authors: Jun Chen, Deyao Zhu, Xiaoqian Shen*

Abstract

Large language models have shown their remarkable capabilities as a general interface for various language-related applications. Motivated by this, we target to build a unified interface for completing many vision-language tasks including image description, visual question answering, and visual grounding, among others. The challenge is to use a single model for performing diverse vision-language tasks effectively with simple multi-modal instructions. Towards this objective, we introduce MiniGPT-v2, a model that can be treated as a unified interface for better handling various vision-language tasks. We propose using unique identifiers for different tasks when training the model. These identifiers enable our model to better distinguish each task instruction effortlessly and also improve the model learning efficiency for each task. After the three-stage training, the experimental results show that MiniGPT-v2 achieves strong performance on many visual question-answering and visual grounding benchmarks compared to other vision-language generalist models. Our model and codes are available at https://minigpt-v2.github.io/